3.1 Design and Implementation of the POWER5™ Microprocessor

Joachim Clabes[1], Joshua Friedrich[1], Mark Sweet[1], Jack DiLullo[1], Sam Chu[1], Donald Plass[2], James Dawson[2], Paul Muench[2], Larry Powell[1], Michael Floyd[1], Balaram Sinharoy[2], Mike Lee[1], Michael Goulet[1], James Wagoner[1], Nicole Schwartz[1], Steve Runyon[1], Gary Gorman[1], Phillip Restle[3], Ronald Kalla[1], Joseph McGill[1], Steve Dodson[1]

[1]IBM Systems Group, Austin, TX
[2]IBM Systems Group, Poughkeepsie, NY
[3]IBM Research, Yorktown Heights, NY

POWER5™ is the next generation of IBM's POWER microprocessors. This design, shown in Fig. 3.1.1, sets a new standard of server performance by incorporating simultaneous multithreading (SMT), an enhanced distributed switch and memory subsystem supporting 1-64w SMP, and extensive RAS support. First pass hardware using IBM's 130nm silicon-on-insulator technology operates above 1.5GHz at 1.3V.

POWER5's dual-threaded SMT [1] creates up to two virtual processors per core, improving execution unit utilization and masking memory latency. Although a simplistic SMT implementation promised ~20% performance improvement, resizing critical micro-architectural resources almost doubles in many cases the SMT performance benefit at a 24% area cost per core.

The two SMT cores interface with an enhanced memory subsystem. The cache hierarchy includes a larger (1.9MB) L2 cache, reduced L3 latency, and a larger (36MB) L3 cache located on a custom DRAM companion chip. The new on-chip main memory controller improves latency and the enhanced interconnect fabric extends SMP scalability. Figure 3.1.2 depicts the microarchitectural changes introduced with POWER5 chip.

POWER5 also greatly enhances the RAS features of POWER4™ by adding the following features:
- dynamic firmware updates;
- full ECC on all interconnects, including address and tag;
- concurrent CEC maintenance;
- additional centralized resource redundancy.

Implementing these mircoachitectural enhancements posed challenges in meeting the chip's frequency, area, power, and thermal targets.

To achieve the target cycle time, the design fully leverages IBM's partially-depleted 130nm SOI process with eight levels of high-performance, copper wiring and FTEOS dielectric. POWER5 also required the redesign of many high performance circuits and the optimization of our integration, timing, and noise-analysis methodologies.
- Restructuring the L1 instruction cache into a 2-way associative array with a small external set-prediction array improved performance and provided cycle time relief.
- Cycle-boundary shifts allowed the performance sensitive 2 cycle access and late-select paths of the 10-way associative L2 cache to meet the cycle-time target.
- Redesigning the data ERAT as a Sum-Addressed-CAM (SACAM) array created a fully associative array that still met the aggressive cycle time.
- Extensive use of an elastic interface design for chip IOs enabled higher frequency system buses with optimum latency. These interfaces stay continuously tuned via periodic self-calibration.
- A semi-automatic tool routed the many cycle-time critical, long wire paths using pre-placed spare buffers under set distance constraints.
- Extensive PowerSPICE simulation identified the most cost effective wiring solutions for the critical core interfaces and SMP fabric bus controllers.
- As shown in Fig. 3.1.3, the clock distribution challenges posed by the larger chip size and additional asynchronous clocking domain were overcome by the use of two on-chip PLLs and an enhanced clock skew minimization tool. [2]
- Supporting both asynchronous and synchronous operation between the processor and memory controller clock domains required additional early and late mode timing analysis.

Although the total transistor count increased to 276M, higher area efficiency offset some of the resulting growth. Unit floor plans were altered to reduce white-space while ensuring wireability. Efficient array and register file redesign also reduced area growth while supporting SMT performance increases. Rather than increasing in size, the L1 data cache improved its performance by becoming 4-way associative. Redesigning the floating point and general purpose register files increased the available register renames with minimal growth.

Further design changes address the AC and DC power concerns stemming from the larger transistor count. POWER5 exploits IBM's triple Vt devices and triple gate oxide process to reduce DC power by more than 40%. 30% of POWER5's transistors are high Vt (mostly in arrays and non-timing-critical logic) while only 0.4% are low Vt (down from more than 7% in POWER4). Using thick oxide decoupling cells eliminates respective gate leakage while sacrificing ~20% capacitance. Dynamic clock gating reduces switching power by more than 25% with no impact on frequency or performance. To minimize di/dt noise caused by simultaneously gating large processor units, POWER5 contains fine-grain gating domains. Detailed analysis of workload simulation, power grid characteristics, and macro-level power ensured adequate decoupling for induced noise. Programmability of all gating events further minimized the functional and noise risks of this new feature. Figure 3.1.4 illustrates the impact of POWER5's power-saving features and shows a circuit concept for clock gating.

POWER5 employs 24 digital temperature sensors to protect the chip from overheating in case of adverse environmental conditions. Each thermal sensor consists of a ring oscillator whose frequency is controlled by a temperature-sensitive current reference and a counter that records the number of oscillations within set time interval. Programmable registers define the maximum allowed temperature on each sensor. When an over-temperature condition occurs, the sensors signal the core's control logic to engage a dual-staged, temperature-reducing response. The first stage reduces average switching and clocking power by rapidly alternating between execution and stall conditions. Once the temperature falls below a reset value, normal operation resumes. If the first thermal response fails to reduce the temperature within an acceptable time period, the second stage will engage and dramatically reduce temperature by prolonged throttling the processor throughput via functions such as fetch, dispatch, or completion. Figure 3.1.5 illustrates this thermal protection mechanism. The exact duration, intensity, and mechanism of each response is fully programmable. Moreover, because the throttling response requires no software or service processor intervention, it provides timely and flexible protection for 1-64w systems while minimizing performance impact. Figure 3.1.6 shows the hardware-measured response of a hot spot's temperature as the chip moves in and out of stage 1 throttling.

Acknowledgments:
The authors thank Carl Anderson, Ravi Arimilli, and the entire POWER5 design team.

References:
[1] R. Kalla, B. Sinharoy, J. Tendler, "A SMT Implementation in POWER5," Hot Chips, Aug. 2003.
[2] P. J. Restle et al, "A Clock Distribution Method for Microprocessors," *IEEE J. Solid-State Circuits*, vol. 36, pp. 792-799, May 2001.

- Technology: 130nm lithography, SOI, Cu
- 276M transistors
- 389 mm^2 die size
- Above 1.5GHz laboratory operation at 1.3V
- 8-way superscalar, 2-way simultaneous multithreaded SMT core
- Enhanced memory subsystem with 1.9MB L2 cache, on-chip L3 directory and memory controller
- Scalable to 64-way SMP
- Extensive RAS support

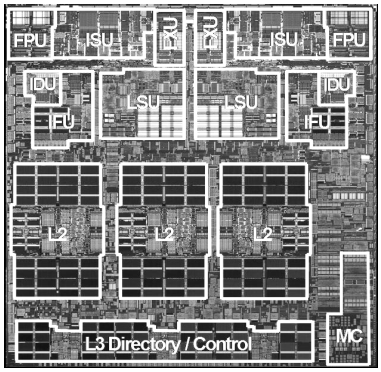
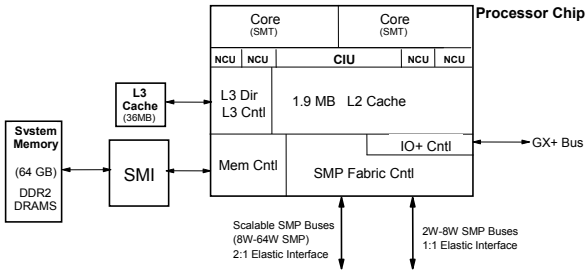
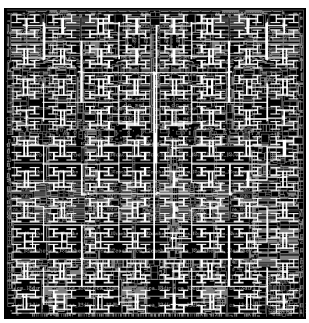


Figure 3.1.1: POWER5 overview.



POWER4 System Features	POWER5 System Features
Two cores	Two dual-threaded SMT cores
1.44 MB L2 cache; 32MB L3 cache	1.9MB L2 cache; 36MB L3 cache
ASIC memory controller chip	On-processor memory controller
32-way SMP scalability	64-way SMP scalability
Mixture of elastic and synchronous interfaces	Elastic interface on all IO which supports greater than 2GHz operation

Figure 3.1.2: System-level view of POWER5.



Main Clock Distribution (91 Buffers):
- 1 full chip buffer
- 1 central chip buffer
- 3 half chip buffers
- 6 quadrant buffers
- 80 sector buffers

Memory Clock Distribution (4 Buffers):
- 1 central chip buffer
- 3 sector buffers

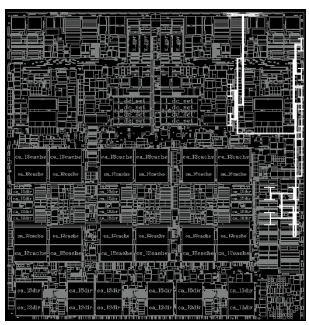


Figure 3.1.3: Dual clock distribution on POWER5.

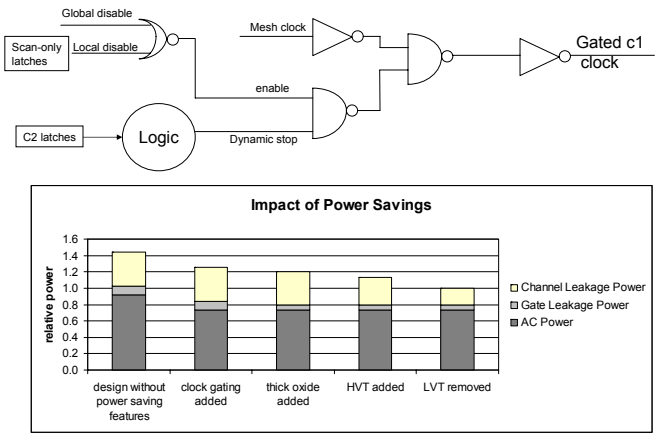


Figure 3.1.4: Clock gating schematic and power savings impact.

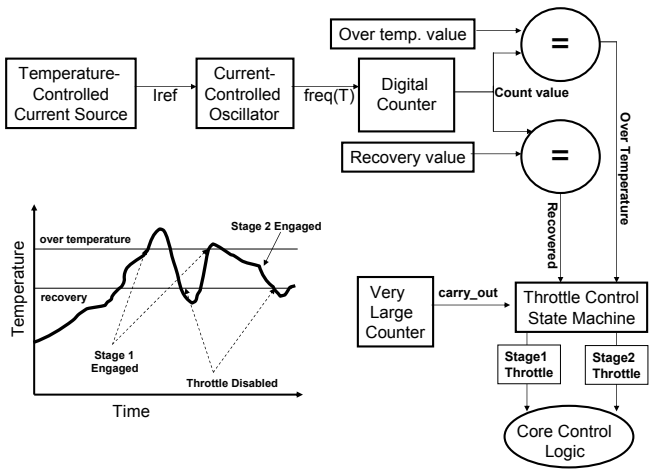


Figure 3.1.5: Thermal control logic and sample thermal response.

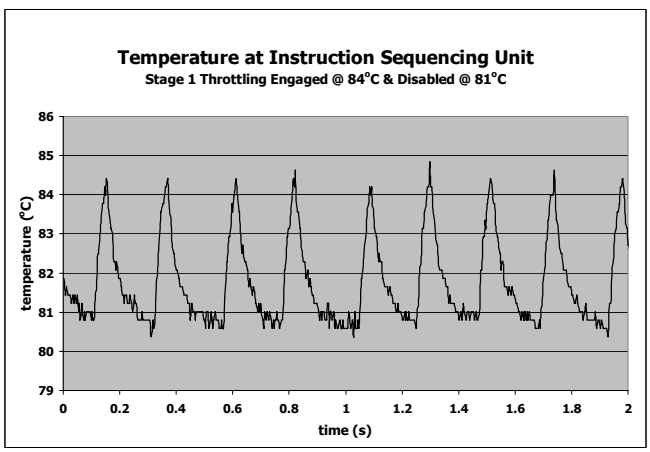


Figure 3.1.6: Temperature response with thermal throttling engaged.

 0-7803-8267-6/$17.00

- Technology: 130nm lithography, SOI, Cu
- 276M transistors
- 389 mm^2 die size
- Above 1.5GHz laboratory operation at 1.3V
- 8-way superscalar, 2-way simultaneous multithreaded SMT core
- Enhanced memory subsystem with 1.9MB L2 cache, on-chip L3 directory and memory controller
- Scalable to 64-way SMP
- Extensive RAS support

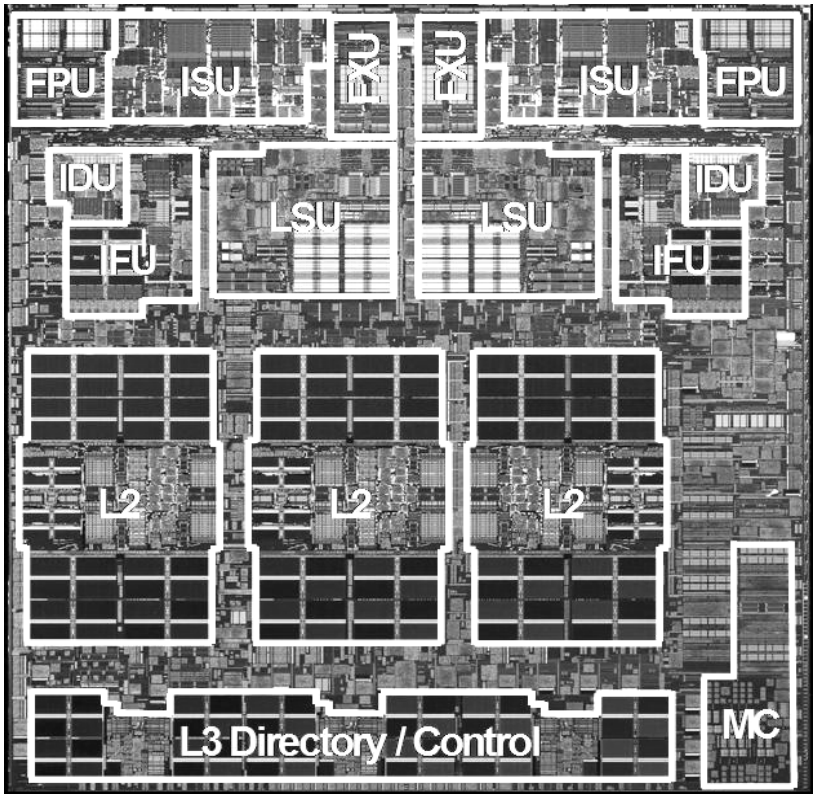
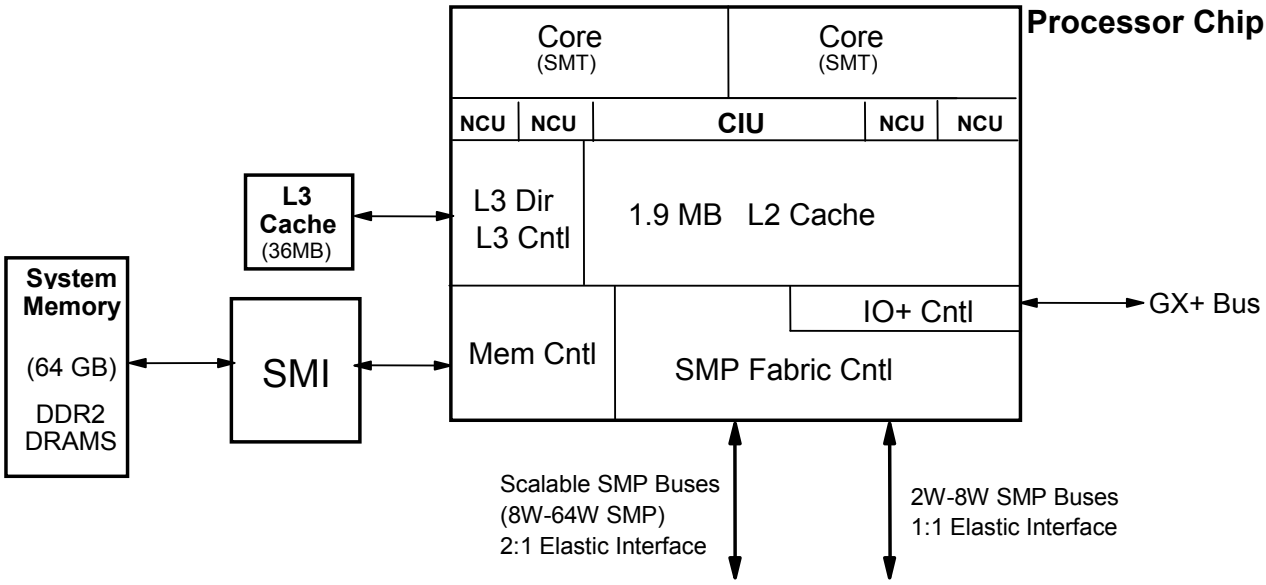


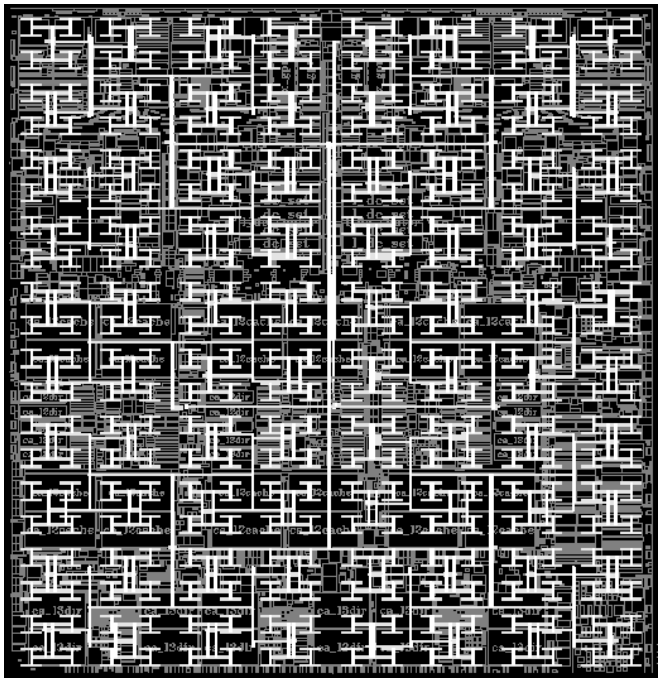
Figure 3.1.1: POWER5 overview.

0-7803-8267-6/04



POWER4 System Features	POWER5 System Features
Two cores	Two dual-threaded SMT cores
1.44 MB L2 cache; 32MB L3 cache	1.9MB L2 cache; 36MB L3 cache
ASIC memory controller chip	On-processor memory controller
32-way SMP scalability	64-way SMP scalability
Mixture of elastic and synchronous interfaces	Elastic interface on all IO which supports greater than 2GHz operation

Figure 3.1.2: System-level view of POWER5.



Main Clock Distribution (91 Buffers):
- 1 full chip buffer
- 1 central chip buffer
- 3 half chip buffers
- 6 quadrant buffers
- 80 sector buffers

Memory Clock Distribution
(4 Buffers):
- 1 central chip buffer
- 3 sector buffers

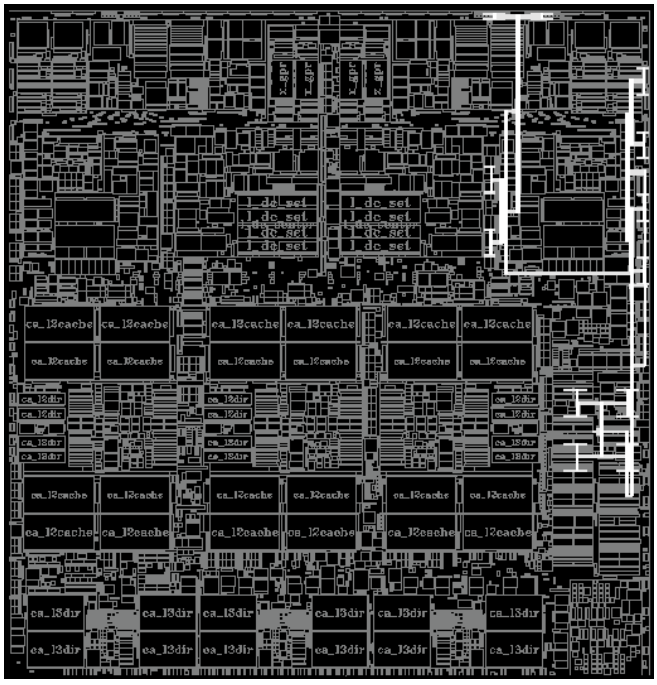
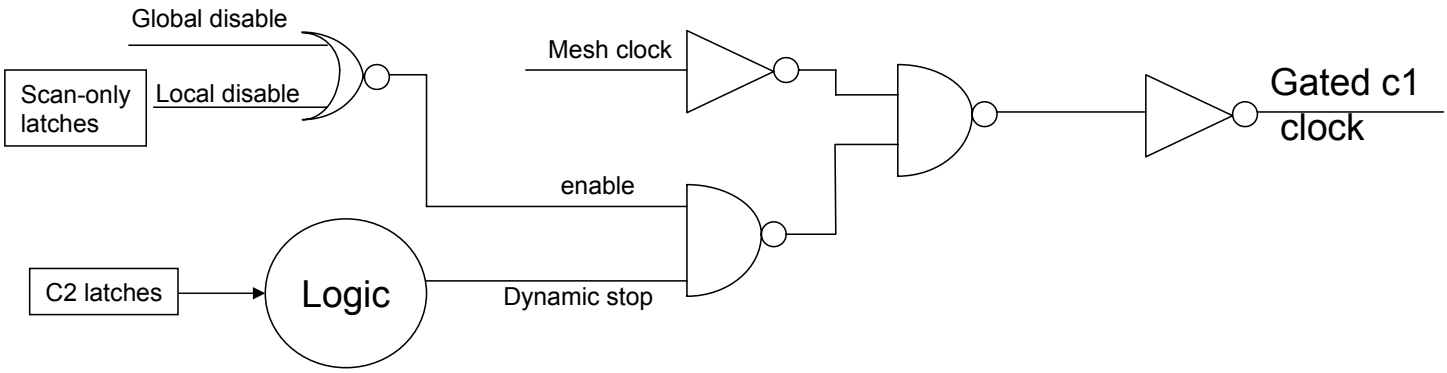


Figure 3.1.3: Dual clock distribution on POWER5.



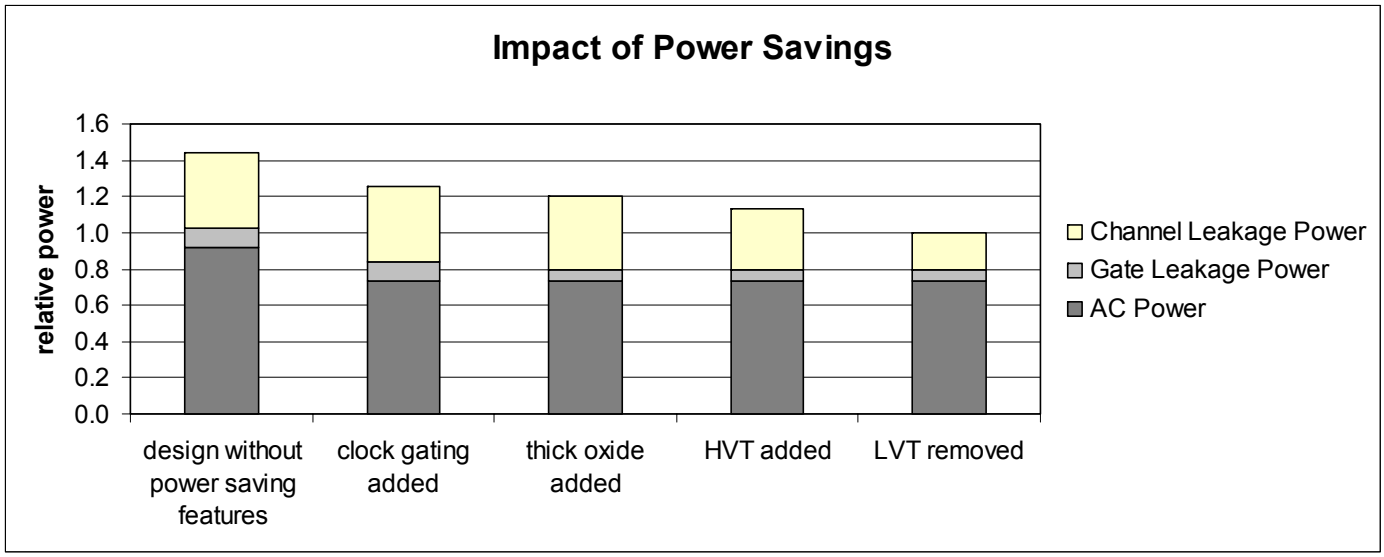


Figure 3.1.4: Clock gating schematic and power savings impact.

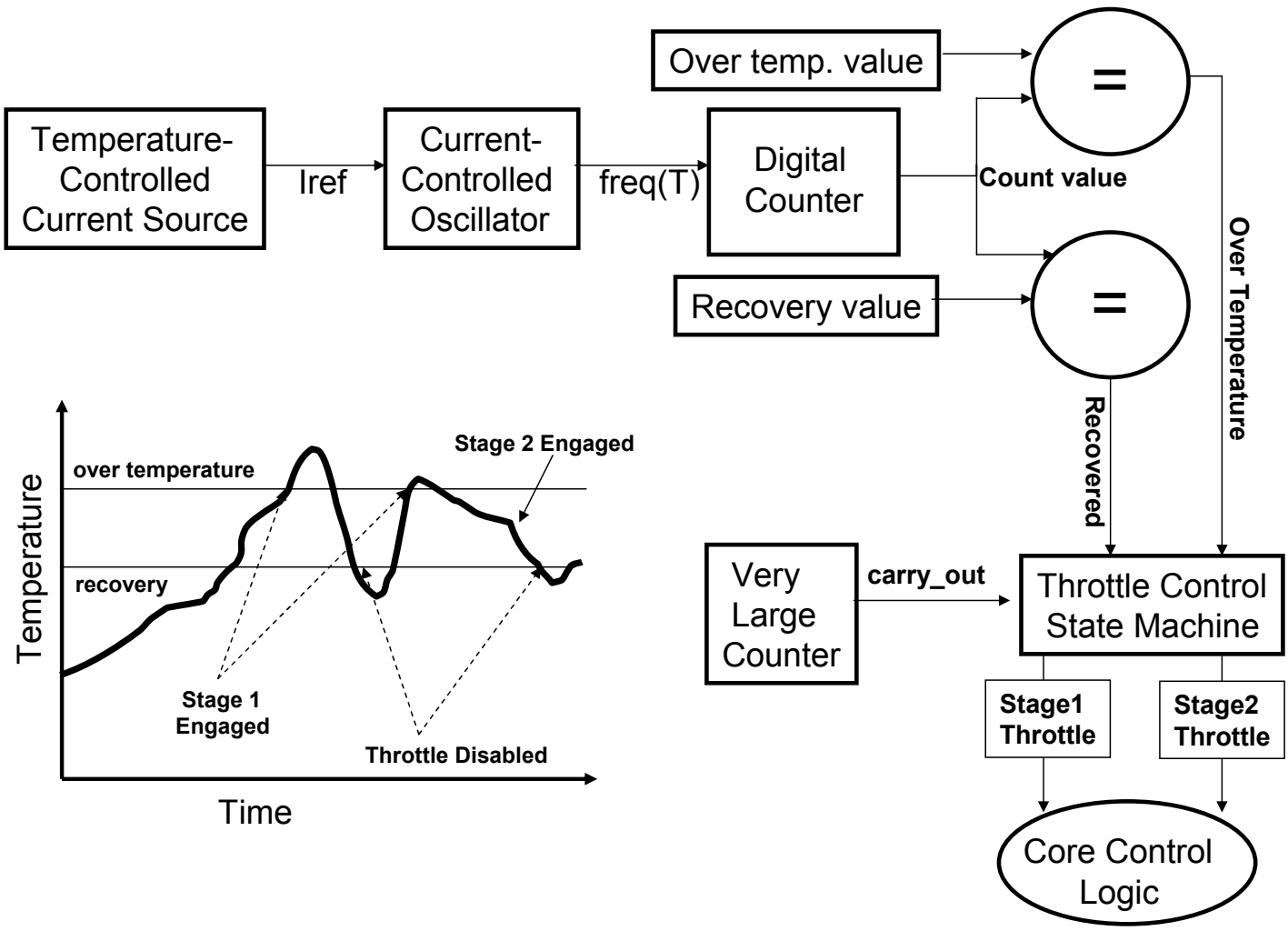


Figure 3.1.5: Thermal control logic and sample thermal response.

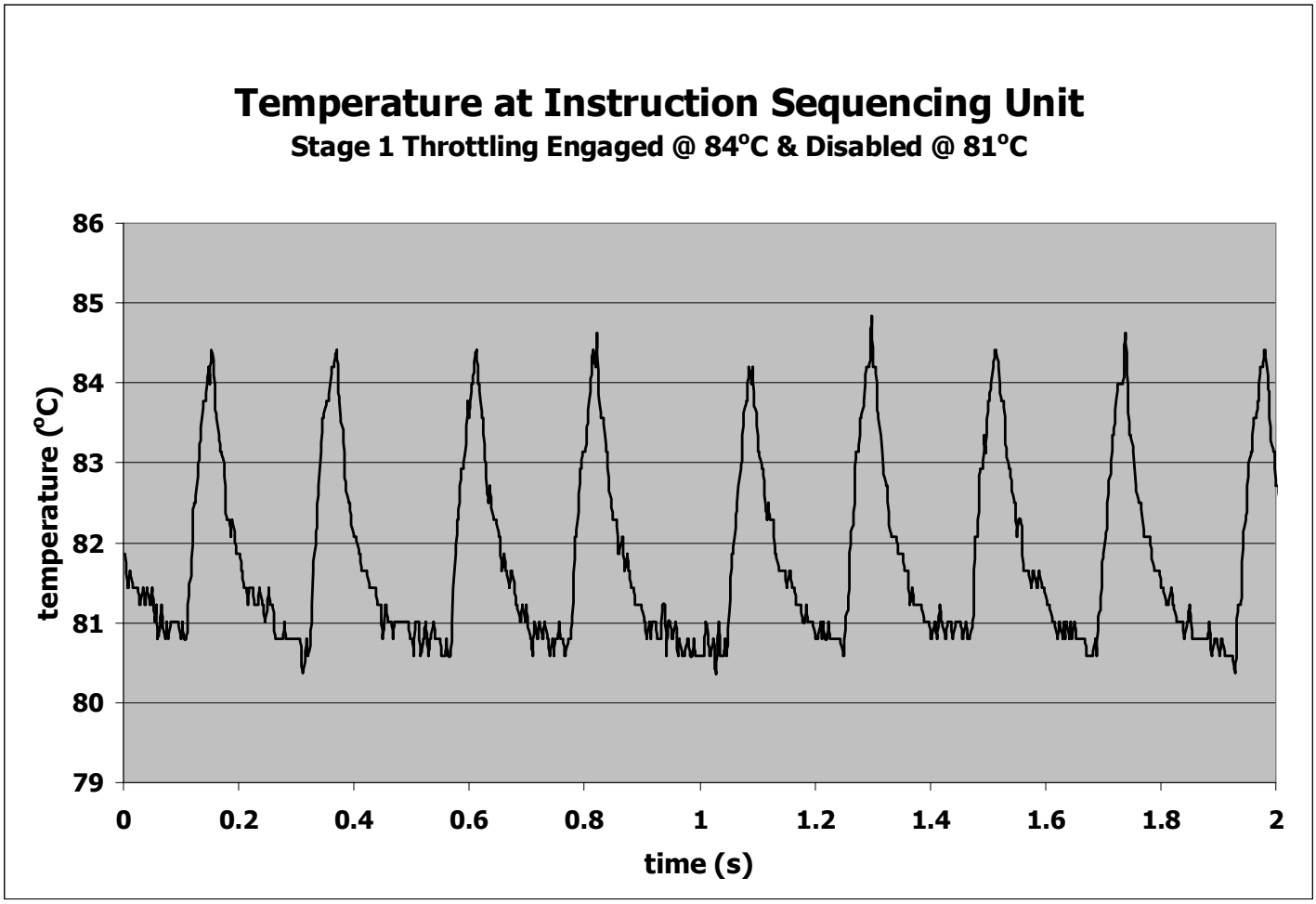


Figure 3.1.6: Temperature response with thermal throttling engaged.

0-7803-8267-6/04